ISSN 1561-5359. Штучний інтелект, 2017, № 3-4

УДК 004.62:004.023

О.Р. Чертов, Д.Ю. Тавров
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Україна
пр. Перемоги, 37, м. Київ, 03056

ЗАБЕЗПЕЧЕННЯ ГРУПОВОЇ АНОНІМНОСТІ ЯК СКЛАДОВА CSID-ПРОЦЕСУ ОБРОБКИ ДАНИХ

O.R. Chertov, D.Y. Tavrov
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine
37, Peremohy av., Kyiv, 03056

PROVIDING GROUP ANONYMITY AS A PART OF CSID DATA PROCESS

У статті задачу забезпечення групової анонімності розглянуто в контексті CSID-процесу обробки даних, виконано порівняльний аналіз описаних у літературі методів її забезпечення. На основі аналізу сформульовано умови, за яких доцільніше використовувати відповідні методи.

Ключові слова: групова анонімність, CSID-процес, міметичний алгоритм, мікрофайл.

In the article, the task of providing group anonymity is discussed in the context of the CSID data process. A comparative study of appropriate methods given in the literature is performed. Based on this study, conditions are formulated for choosing methods that fit each particular case.

Key words: group anonymity, CSID data process, memetic algorithm, microfile.

Вступ

У світі спостерігається невпинне зростання обсягів цифрових даних, значна частина яких містить конфіденційну інформацію про особу чи групу осіб. Упередження витоку такої інформації можна забезпечити одним із двох способів [1]: трансформувати даних для зниження ризику розкриття інформації або фізично обмежити доступ до даних. Оскільки в умовах інформаційного суспільства значна частка даних публікується у відкритому доступі, перший підхід часто є єдиним можливим.

Як правило, інтереси респондентів та кінцевих користувачів даних прямо суперечать один одному, оскільки перші зацікавлені в максимальному захисті чутливої інформації про себе, а другі — у здобутті якомога повніших первинних неагрегованих даних (*мікроданих*). Організації, які забезпечують публічний доступ до (мікро)даних із одночасним захистом від порушення приватності осіб та груп осіб, називають *організаціями-розпорядниками даних* (data stewardship organizations) [2]. Прикладами таких організацій є державні статистичні служби (наприклад, Державна служба статистики України), міжнаціональні статистичні установи (наприклад, Статистичний офіс Європейського Співтовариства), медичні заклади, архіви тощо.

Процес обробки даних, який здійснюють організації-розпорядники, можна розділити [2] на чотири підпроцеси — збір (capture), зберігання (storage), інтеграція (integration) та розповсюдження (dissemination). Такий процес називають CSID-процесом за першими літерами відповідних англійських термінів. Ці підпроцеси передбачають:

– збір даних шляхом спостережень, переписів населення чи опитувань;
– збереження даних у великих обсягах у доступному електронному форматі;
– інтеграцію даних між різними базами даних;
– розповсюдження даних, результатом якого є деякий інформаційний продукт, наприклад, вихідні таблиці чи файли мікроданих.

Виділяють дві моделі розповсюдження даних [3]:

- модель із недовірою, за якої організація-розпорядник не користується довірою респондентів та може порушити приватність зібраних даних;
- модель із довірою, за якої організація-розпорядник користується довірою респондентів, тобто збір даних відбувається без застосування додаткових методів захисту. При цьому довіра не поширюється на потенційного кінцевого користувача інформаційного продукту, тому розповсюдження даних повинно передбачати додатковий їх захист.

У даній роботі розглядатимемо тільки модель із довірою.

Маскувати дані можна на трьох різних стадіях CSID-процесу [4]: на етапі збору даних (шляхом рандомізованих опитувань), на етапі розповсюдження даних або на етапі подання результатів їх аналізу. У даній роботі нас цікавить тільки друга можливість. На етапі розповсюдження організація-розпорядник даних, серед іншого, повинна забезпечити анонімність публікованих даних. У силу надзвичайної практичної значущості цієї задачі систематизації методів її розв’язання присвячено дану роботу.

Постановка проблеми

Під анонімністю об’єкта розуміють неможливість однозначного характеризувати його у множині певних об’єктів [5]. Анонімність буває двох видів:

- індивідуальна анонімність, яка стосується інформації про окремого респондента;
- групова анонімність, яка стосується розподілу інформації про групу респондентів.

Забезпечення анонімності передбачає виконання таких етапів [2]:

- аналіз ризику порушення анонімності;
- модифікація даних для зменшення ризику;
- аналіз впливу застосованих модифікацій на корисність даних.

У даній роботі розглядатимемо тільки забезпечення групової анонімності даних. У літературі описано низку відповідних методів, але не вказано умов, за яких використання того чи іншого методу є найбільш доцільним. Тому в даній роботі ставиться задача на основі критичного аналізу літератури з групової анонімності систематизувати існуючі методи її забезпечення та встановити, які методи найбільш ефективно використовувати і за яких умов.

Аналіз останніх досліджень і публікацій

Методи індивідуальної анонімізації можна розділити на пертурбативні (модифікують записи з набору даних) та непертурбативні (анонімізують дані, явно не спотворюючи їх). До пертурбативних методів належать додавання до даних шуму з метою ускладнення ідентифікації записів [6], методи досягнення *k*-анонімності [7], обмін значень певних атрибутів між різними записами [8] тощо. До непертурбативних методів належать перекодування та огрублення даних [9] та ін.

В останні роки розроблено додаткові методи забезпечення індивідуальної анонімності:

- алгоритм захисту приватності [10], який передбачає переробку даних з метою унеможливлення віднаходження класифікаційних правил, які можуть призводити до витоку чутливої інформації;
- метод на основі кластеризації [11] у випадку декількох чутливих атрибутів даних;
- методи забезпечення індивідуальної анонімності в соціальних мережах [12].

Огляд інших новітніх методів наведено в [13].

Уведемо деякі позначення. Називатимемо *мікрофайлом* $\mathbf{M}$ таблицю з мікроданими, у якій рядки (*записи*) $\mathbf{r}^{(i)}$, $i = 1,...,\rho$, відповідають респондентам, а

стовпці w_j, $j=1,...,\eta$, — атрибутам. *Сутнісними атрибутами* w_{v_j}, $j=1,...,l$, називатимемо атрибути, значення яких дають змогу ідентифікувати респондента як належного деякій *групі* (підмножині записів). *Сутнісними записами* $\mathbf{r}_v^{(i)}$, $i=1,...,\rho_v$, називатимемо записи, значення сутнісних атрибутів яких належать декартовому добутку $\mathbf{V}=\mathbf{w}_{v_1}\times\ldots\times\mathbf{w}_{v_l}$. Називатимемо *параметризуючим* атрибут w_p, $p\neq v_j$ $\forall j$, значення якого $\mathbf{P}=\left\{P_i \mid P_i\in\mathbf{w}_p\right\}$, $i=1,...,l_p$, — *параметризуючі значення* — визначають розподіл інформації про сутнісні записи. Називатимемо *параметричними підмікрофайлами* підмножини записів $\mathbf{M}_1,\ldots,\mathbf{M}_{l_p}$, у кожній із яких записи мають однакове параметричне значення. Кількість записів в *і*-ому підмікрофайлі позначатимемо через ρ_i.

Розподіл даних про групу за значеннями параметризуючого атрибуту називатимемо *цільовим сигналом* $\boldsymbol{\theta}=\left(\theta_1,\theta_2,...,\theta_{l_p}\right)$. У роботі розглядатимемо тільки один його різновид — *кількісний сигнал* $\mathbf{q}=\left(q_1,q_2,...,q_{l_p}\right)$, де q_k — кількість сутнісних записів в $\mathbf{M}_k$. Вибір інших описаних у літературі типів цільових сигналів не впливає на одержані в статті висновки.

Ризик порушення групової анонімності несуть *викиди* кількісного сигналу *вгору*, які відповідають аномальним скупченням респондентів з однаковими характеристиками (сутнісними комбінаціями значень) в одному параметричному підмікрофайлі. Множину індексів $\mathbf{q}$, які відповідають викидам, позначатимемо через $OUT(\mathbf{q})$.

Як відомо [14], вилучення сутнісних атрибутів (наприклад, «Військова служба») із мікрофайлу не забезпечує групової анонімності в загальному випадку, оскільки [15] існує можливість сформувати модель групи респондентів мікрофайлу на основі значень решти атрибутів мікрофайлу, вилучення яких неприпустиме (наприклад, «Вік» чи «Стать»), яка дає можливість побудувати розподіл, викиди якого відповідають викидам кількісного сигналу.

У літературі виділяють [16] два основні підходи до забезпечення групової анонімності даних: одноетапний та двоетапний. У двоетапних методах спочатку здійснюють модифікацію кількісного сигналу для маскування викидів, а потім модифікують дані мікрофайлу, щоб по них можна було побудувати модифікований сигнал. Такий підхід простий у застосуванні (можна задати довільний вигляд модифікованого сигналу), проте на практиці призводить до внесення в дані спотворень великого обсягу (оскільки під час визначення вигляду модифікованого сигналу не враховуються інші атрибути, окрім сутнісних).

В одноетапних методах цільовий сигнал та мікрофайл модифікуються одночасно шляхом послідовного обміну між різними підмікрофайлами схожих респондентів, один із яких належить групі, а інший — ні. У цьому випадку з’являється можливість явно сформулювати критерій оптимальності розв’язку ЗЗГА, який враховує як обсяг внесених у мікрофайл спотворень, так і рівень маскування викидів у початковому сигналі.

Мета дослідження

Метою даної роботи є систематизація існуючих методів забезпечення групової анонімності в рамках підпроцесу розповсюдження даних CSID-процесу та встановлення умов, за яких найбільш застосовні відповідні методи.

Формалізація задачі забезпечення групової анонімності

Розв'язання ЗЗГА передбачає знаходження послідовності пар записів мікрофайлу $\mathbf{S} = \left(\left(\mathbf{r}^{(i_1)}, \mathbf{r}^{(j_1)} \right), \ldots, \left(\mathbf{r}^{(i_Q)}, \mathbf{r}^{(j_Q)} \right) \right)$, де i_k, j_k, $k = \overline{1, Q}$ — індекси записів мікрофайлу, які характеризуються такими властивостями:

– записи в кожній парі належать різним підмікрофайлам;
– у кожній парі тільки один запис є сутнісним;
– кількісний сигнал, збудований за модифікованим мікрофайлом, утвореним після виконання обмінів із $\mathbf{S}$, не містить початкових викидів (або принаймні значної їх частини);
– записи в парі повинну бути в деякому сенсі схожі між собою, щоб обсяг спотворень, унесених у дані, був якомога менший.

У літературі як міру схожості записів використовують [16] *визначальну метрику*:

$$\mathrm{InfM}(\mathbf{r}, \mathbf{r}^*) = \sum_{k=1}^{n_{\text{пор}}} \omega_k \left(\frac{r_{I_k} - r_{I_k}^*}{r_{I_k} + r_{I_k}^*} \right)^2 + \sum_{l=1}^{n_{\text{кат}}} \gamma_l \chi^2 \left(r_{J_l}, r_{J_l}^* \right), \quad (1)$$

де I_k (J_l) — k-ий порядковий (l-ий категорійний) *визначальний атрибут* (атрибут, розподіл значень якого становить інтерес для дослідників), $\chi(v_1, v_2)$ дорівнює деякому числу χ_1, якщо v_1 та v_2 належать одній категорії, та χ_2 — у протилежному випадку, ω_k та γ_l — невід'ємні вагові коефіцієнти (що важливіший атрибут, то більше значення відповідної ваги).

У [17] показано, що ЗЗГА можна звести до задачі пошуку потоку мінімальної вартості в мережі [18], архітектура якої безпосередньо визначається вибором сутнісних та параметризуючих значень, а також значеннями модифікованого кількісного сигналу $\mathbf{q}^*$. Позначмо через $G = (N, A)$ орієнтовану мережу, визначену множиною N із n вузлів та множиною A з m орієнтованих дуг. Кожна дуга характеризується вартістю c_{ij} та пропускною здатністю u_{ij}. Кожний вузол i асоційовано з деяким числом $b(i)$, яке можна інтерпретувати як його пропозицію (якщо $b(i) > 0$) або попит (якщо $b(i) < 0$). Задачу пошуку максимального потоку мінімальної вартості x можна сформулювати так:

$$\min \sum_{(i, j \in A)} c_{ij} x_{ij},$$

$$\sum_{j | (i, j \in A)} x_{ij} - \sum_{j | (j, i \in A)} x_{ji} = b(i) \quad \forall i \in N,$$

$$0 \le x_{ij} \le u_{ij} \quad \forall (i, j) \in A,$$

$$\sum_{i=1}^{n} b(i) = 0.$$

Архітектура мережі у випадку ЗЗГА має такі особливості (рис. 1):

– $N = N_1 \cup N_2 \cup N_3 \cup N_4$, $N_i \cap N_j = \varnothing \ \forall i \ne j$;
– вузли $N_l^k \in N_l$, $l = \overline{1, 4}$, відповідають параметричним підмікрофайлам $\mathbf{M}_k$;

– N_2 (N_3) містить l_p підмножин, що не перетинаються. Вузли $N_2^{(k,i)} \in N_2^{(k)}$ ($N_3^{(k,i)} \in N_3^{(k)}$) відповідають сутнісним (несутнісним) записам $\mathbf{M}_k$, $k = 1,...,l_p$, $i = 1,...,q_k$ ($i = 1,...,\rho_k - q_k$);
– $A = A_{N_1N_2} \cup A_{N_2N_3} \cup A_{N_3N_4}$, $A_{N_iN_j} \cap A_{N_kN_l} = \varnothing \ \forall i \neq k, j \neq l$;
– дуги в $A_{N_iN_j}$ з'єднують кожний вузол із N_i з кожним вузлом із N_j;
– пропозиції вузлів із N_1 дорівнюють $b\left(N_1^k\right) = \begin{cases} \delta_k, \delta_i > 0 \\ 0, \ \delta_i \leq 0 \end{cases}$, де $\delta_k = q_k - q_k^*$;
– попити вузлів із N_4 дорівнюють $b\left(N_4^k\right) = \begin{cases} \delta_k, \delta_i > 0 \\ 0, \ \delta_i \leq 0 \end{cases}$, де $\delta_k = q_k - q_k^*$;
– пропозиції та попити вузлів в N_2 та N_3 дорівнюють 0;
– $u_{ij} = 1 \ \forall i, j$; $c_{ij} = 0 \ \forall i, j$, якщо $c_{ij} \in A_{N_1N_2}$ або $c_{ij} \in A_{N_3N_4}$; вартість дуги $\left(N_2^{(k,i)}, N_3^{(l,j)}\right) \in A_{N_2N_3}$, $k = 1,...,l_p$, $l = 1,...,l_p$, $k \neq l$, $i = 1,...,q_k$, $i = 1,...,\rho_l - q_l$, дорівнює значенню (1), обчисленому для відповідної пари записів.

ЗЗГА можна сформулювати так: знайти послідовність $\mathbf{S}$, яка задовольняє умови

$$\begin{gathered} \frac{\left|OUT(\mathbf{q}) \cap OUT\left(\mathbf{q}^*(\mathbf{S})\right)\right|}{\left|OUT(\mathbf{q})\right|} \leq K_{out}, \\ \sum_{k=1}^{Q} \mathrm{InfM}\left(\mathbf{r}^{(i_k)}, \mathbf{r}^{(j_k)}\right) \leq K_{dist} \cdot C_{\max}, \end{gathered} \tag{2}$$

де $\mathbf{q}^*(\mathbf{S})$ — модифікований сигнал після послідовних попарних обмінів записів із $\mathbf{S}$, K_{out} — *поріг чутливості*, K_{dist} — *поріг спотворень*, $C_{\max}$ — найбільше можливе сумарне значення визначальної метрики (1), яку можна обчислити для розв'язуваної ЗЗГА.

Двоетапні методи розв'язання задачі забезпечення групової анонімності

Історично першими було запропоновано двоетапні методи забезпечення групової анонімності даних. Етапи підпроцесу розповсюдження даних у цьому випадку можна інтерпретувати таким чином:
– аналіз ризику порушення анонімності: виявлення викидів у кількісному сигналі. Якщо їх можна виявити візуально чи за допомогою спеціальних методів, вважається, що ризик порушення анонімності існує;
– модифікація даних для зменшення ризику: у літературі описано три методи, які передбачають модифікацію кількісного сигналу для маскування викидів, які розглянемо нижче, а також метод модифікації мікрофайлу для приведення його у відповідність до модифікованого кількісного сигналу на основі розв'язання відповідної задачі пошуку потоку в мережі;
– аналіз впливу застосованих модифікацій на корисність даних: для цього в кожному з методів запропоновано свій спосіб оцінки якості одержуваного модифікованого сигналу, а для оцінки загального обсягу внесених спотворень використовується метрика (1).

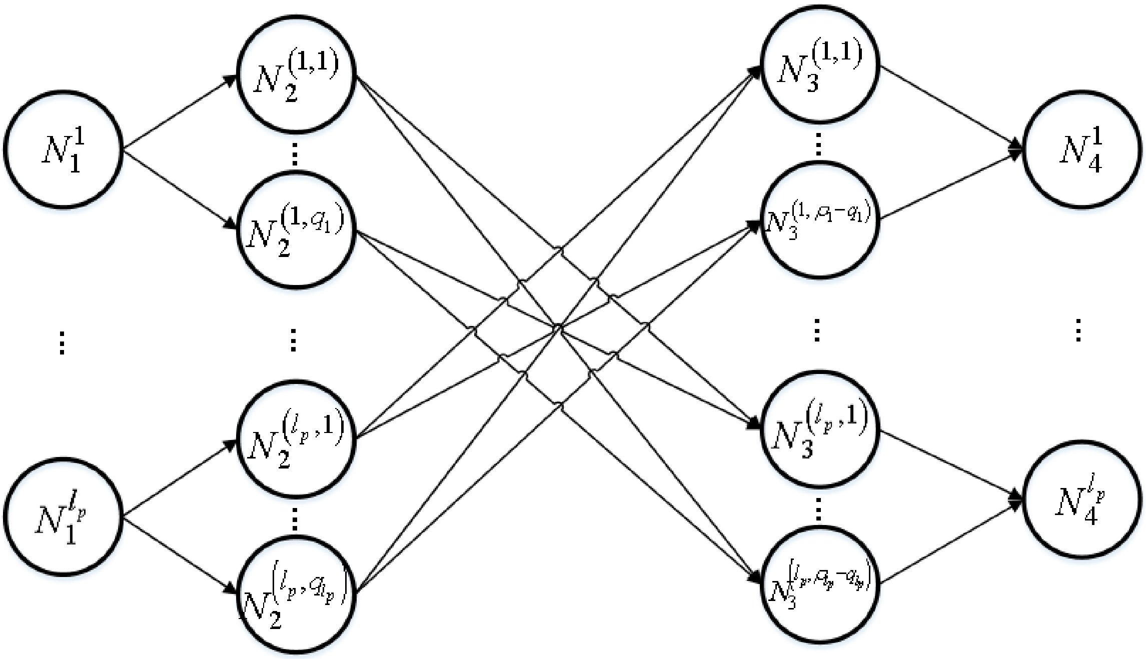


Рис. 1. Архітектура мережі для задачі забезпечення групової анонімності

Різні методи модифікації кількісного сигналу дають змогу зберегти різні властивості сигналу, маскуючи при цьому його викиди. Зокрема, *метод нормалізації* [19] дає змогу зберегти середнє та середньоквадратичне відхилення сигналу. Для цього потрібно:

– довільним чином модифікувати $\mathbf{q}$ та одержати сигнал $\hat{\mathbf{q}}$, у якому масковано викиди;
– нормалізувати $\hat{\mathbf{q}}$:

$$\mathbf{q}^* = \left(\hat{\mathbf{q}} + \frac{\sigma^*}{\hat{\sigma}} \cdot \hat{\mu} - \mu^*\right) \cdot \frac{\hat{\sigma}}{\sigma^*},$$

де $\hat{\mu} = \sum_{i=1}^{l_p} \hat{q}_i \big/ l_p$, $\mu^* = \sum_{i=1}^{l_p} q_i^* \big/ l_p$, $\hat{\sigma} = \sqrt{\sum_{i=1}^{l_p} \left(\hat{q}_i - \hat{\mu}\right)^2 \big/ \left(l_p - 1\right)}$, $\sigma^* = \sqrt{\sum_{i=1}^{l_p} \left(q_i^* - \mu^*\right)^2 \big/ \left(l_p - 1\right)}$.

Збереження тільки статистичних моментів часто є недостатнім, тому в літературі запропоновано інші методи модифікації кількісних сигналів.

Метод на основі вейвлет-перетворень [20] дає змогу зберегти високочастотні особливості сигналу. Основна ідея методу полягає в тому, що сигнал можна подати як

$$\mathbf{q} = \mathbf{A}_k + \sum_{i=1}^{k} \mathbf{D}_i,$$

де $\mathbf{A}_k$ — *апроксимуюча складова* сигналу (його згладження), $\mathbf{D}_i$ — *деталізуючі складові* різних рівнів (високочастотні коливання в сигналі різних частот).

Для маскування викидів сигналу потрібно модифікувати $\mathbf{A}_k$, а для збереження його високочастотних особливостей — залишити незмінними $\mathbf{D}_i$ (чи змінити їх пропорційно). У результаті модифікації, описаної в [20], можна одержати модифіковану складову $\breve{\mathbf{A}}_k$, у якій масковано потрібні викиди, та сигнал $\breve{\mathbf{q}} = \breve{\mathbf{A}}_k + \sum_{i=1}^{k} \mathbf{D}_i$. Якщо деякі з елементів цього сигналу будуть від'ємними, потрібно додати відповідно велике число γ. Для збереження загального числа сутнісних записів потрібно домножити одержуваний сигнал на відповідний коефіцієнт:

$$\mathbf{q}^* = \left(\breve{\mathbf{q}} + \gamma\right) \cdot \left(\sum_{k=1}^{l_p} q_k\right) \Big/ \left(\sum_{k=1}^{l_p} \left(\breve{q}_k + \gamma\right)\right).$$

Третій описаний у літературі метод модифікації кількісного сигналу базується на сингулярно-спектральному аналізі, основна ідея якого полягає [21] в розкладенні початкового ряду в суму його трендової, періодичної та шумової складових. Для цього сигнал спочатку перетворюють у траєкторну матрицю:

$$\mathbf{X} = \begin{pmatrix} q_1 & q_2 & q_3 & \cdots & q_K \\ q_2 & q_3 & q_4 & \cdots & q_{K+1} \\ q_3 & q_4 & q_5 & \cdots & q_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_L & q_{L+1} & q_{L+2} & \cdots & q_{l_p} \end{pmatrix},$$

де L — довжина вікна, $1 < L < l_p$, K — кількість векторів вкладення, $K = l_p - L + 1$.

Відповідну матрицю можна подати у вигляді сингулярного розкладення

$$\mathbf{X} = \mathbf{X}_1 + \ldots + \mathbf{X}_d ,$$

де $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$, λ_i — власні числа матриці $\mathbf{X}\mathbf{X}^T$, узяті в незростаючому порядку, U_i — ортонормовані власні векторі, що відповідають цим числам, $d = \max\{i : \lambda_i > 0\}$, $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$, $i = 1,...,d$.

Множину індексів $\{1,...,d\}$ можна розбити на групи, які відповідають власним числам великої амплітуди, парним власним числам та власним числам, близьким до нуля. Якщо позначити деяку підмножину індексів через $I = \{i_1,...,i_k\}$, а відповідні матриці — через $\mathbf{X}_I = \mathbf{X}_{i_1} + ... + \mathbf{X}_{i_k}$, то траєкторну матрицю можна подати у вигляді суми

$$\mathbf{X} = \mathbf{X}_{I_1} + \ldots + \mathbf{X}_{I_k} ,$$

де кожна матриця відповідає трендовій (власні числа великої амплітуди), періодичним (парні власні числа) чи шумовій (власні числа, близькі до нуля) складовим сигналу. Переведення кожної з відповідних матриць в одновимірний сигнал описано в [21].

Маскувати викиди кількісного сигналу можна шляхом модифікації його трендової складової. Для збереження корисних властивостей сигналу в цьому випадку потрібно не змінювати його періодичних складових.

Таким чином, можна зробити підсумок, що використання кожного з описаних вище методів випливає з потреби в збереженні тих чи інших особливостей кількісного сигналу:

– якщо є потреба в збереженні середнього та середньоквадратичного відхилення кількісного сигналу, потрібно застосовувати метод нормалізації;
– якщо є потреба в збереженні високочастотних особливостей кількісного сигналу, потрібно застосовувати метод на основі вейвлет-перетворень:
– якщо є потреба в збереженні періодичних складових кількісного сигналу, потрібно застосовувати метод на основі сингулярно-спектрального аналізу.

Одноетапні методи розв'язання задачі забезпечення групової анонімності

Двоетапні методи забезпечення групової анонімності мають спільну ваду: обсяг унесених спотворень визначається вибором модифікованого сигналу, але в усіх

методах його сигналу здійснюється довільно, що на практиці призводить до значних обсягів унесених спотворень. В одноетапних методах розв'язання ЗЗГА кількісний сигнал модифікується з метою маскування викидів з *одночасною* модифікацією даних. Фактично, на третьому етапі підпроцесу розповсюдження даних якість модифікованого кількісного сигналу не є ключовою, а основною стає задача мінімізації загального обсягу спотворень у розумінні (1).

Таким чином, стає можливим підбір модифікованого кількісного сигналу, який відповідає архітектурі мережі з потоком мінімальної вартості. Оскільки підбір такого сигналу є в загальному випадку переборною задачею, та враховуючи той факт, що на практиці анонімізовані дані належать до категорії даних великого обсягу, у більшості випадків недоцільно шукати оптимальний розв'язок ЗЗГА. Часто допустимі розв'язки дають змогу надійно забезпечити анонімність, уносячи при цьому в мікрофайл незначні спотворення.

Модифікуймо постановку ЗЗГА. Будемо шукати послідовність обмінів записів $\mathbf{S}$, яка задовольняє такі умови:

$$\mu\left(q_1^*(\mathbf{S}),\ldots,q_{l_p}^*(\mathbf{S})\right) \geq \alpha_{comp},$$
$$\frac{\left|OUT(\mathbf{q}) \cap OUT\left(\mathbf{q}^*(\mathbf{S})\right)\right|}{\left|OUT(\mathbf{q})\right|} \leq K_{out}, \qquad (3)$$
$$\sum_{k=1}^{Q} \mathrm{InfM}\left(\mathbf{r}^{(i_k)},\mathbf{r}^{(j_k)}\right) \leq K_{dist} \cdot C_{\max},$$

де $\mu\left(q_1^*(\mathbf{S}),\ldots,q_{l_p}^*(\mathbf{S})\right)$ — ступінь сумісності $\mathbf{q}^*$ із обмеженнями, які накладають на значення сигналу $\mathbf{q}$, що відповідають викидам, α_{comp} — *поріг сумісності*; як правило, $\alpha_{comp} > 0,5$.

Обмеження з (3) визначає експерт для кожного значення кількісного сигналу, яке відповідає викиду. Кожне таке обмеження є функцією $\mu_i(x)$ з такими властивостями:

- дорівнює 0 для $x \geq q_j$ (викид не може збільшуватися);
- дорівнює 1 для $x \leq \varepsilon_j$, де ε_j — *порогове значення*, нижче якого повинно спуститися відповідне значення сигналу;
- монотонно спадає до 0, коли $\varepsilon_j \leq x \leq q_j$.

У літературі описано [14] міметичний алгоритм розв'язання поставленої таким чином ЗЗГА. Міметичний алгоритм — це [22] еволюційний алгоритм із додаванням локального пошуку. У міметичному алгоритмі для розв'язання ЗЗГА використовується популяція матриць U розмірності $Q \times 4$, де кожний рядок однозначно задає пару записів для обміну:

- елемент першого стовпця — це індекс підмікрофайлу, із якого потрібно вилучити запис;
- елемент другого стовпця — це індекс запису в рамках відповідного підмікрофайлу;
- елемент третього стовпця — це індекс підмікрофайлу, до якого потрібно додати запис;
- елемент четвертого стовпця — це індекс запису в рамках відповідного підмікрофайлу.

Кожний запис може входити в U тільки один раз. Індекс підмікрофайлу i, $i=1,...,l_p$, не може зустрічатися в 1-ому стовпці більше від q_i разів, а в 3-ому — більше від $(\rho_i - q_i)$ разів.

Пристосованість кожної особини в популяції визначається функцією

$$f(U)=\Upsilon(U)\cdot\Phi(U)\cdot\Psi(U),$$

де $\Upsilon(U)$ відповідає якості розв'язку з погляду мінімізації спотворень, $\Phi(U)$ відповідає якості розв'язку з погляду маскування викидів (добуток відповідних $\mu_i(x)$), $\Psi(U)$ — штрафний терм для упередження необмеженого зростання особинах.

Міметичний алгоритм передбачає виконання такої послідовності кроків:

1. Випадковим чином згенерувати популяцію $P=\{U_i\}$ з μ особин, $i=1,...,\mu$.
2. Застосувати оператор локального пошуку $S(U_i)$ $i=1,...,\mu$.
3. Обчислити значення функції пристосованості (3) для кожної особини.
4. Якщо виконується умова завершення, зупинити алгоритм.
5. Вибрати λ пар батьківських особин; помістити їх у множину P'.
6. Застосувати оператор рекомбінації $R(U_{i_1},U_{i_2})$ до кожної пари особин $\langle U_{i_1},U_{i_2}\rangle$ з P', $i_1=1,...,\lambda$, $i_2=1,...,\lambda$, $i_1 \neq i_2$; помістити нащадків у множину P''.
7. Застосувати оператор мутації $M(U_j)=(M_4 \circ M_3 \circ M_2 \circ M_1)(U_j)$ $\forall U_j \in P''$, $j=1,...,\lambda$, де кожний оператор M_k, $k=1,...,4$ діє окремо на відповідний стовпець U_j.
8. Застосувати $S(U_i)$ до кожної особини з P''.
9. Обчислити значення функції пристосованості (3) кожної особини з P''.
10. Вибрати μ найпристосованіших особин із $P \cup P''$; додати їх у P замість поточних.
11. Перейти на крок 3.

Початкова популяція формується шляхом випадкового генерування особин із різною кількістю рядків. Імовірності для генерування елементів першого стовпця пропорційні значенням відповідних елементів $\mathbf{q}$, елементів третього — пропорційні розмірам відповідних підмікрофайлів. Завершення роботи алгоритму можна здійснювати після генерації наперед визначеної кількості поколінь. Вибір решти параметрів алгоритму залежить від особливостей кожної конкретної ЗЗГА.

Порівняння двоетапних та одноетапних методів розв'язання ЗЗГА

На основі аналізу методів розв'язання ЗЗГА можна зробити такі висновки:

– двоетапні методи дають змогу не тільки маскувати викиди кількісного сигналу шляхом його модифікації, а й зберегти деякі його характеристики (моменти, високочастотні складові, періодичні компоненти);
– одноетапні методи в загальному випадку дають змогу модифікувати дані, унісши спотворення меншого обсягу, але корисна інформація, яка може міститися в структурі кількісного сигналу, втрачається.

Систематизацію відповідних спостережень наведено на UML-діаграмі діяльності працівника організації-розпорядника даних, в обов'язки якого входить забезпечення групової анонімності даних (рис. 2).

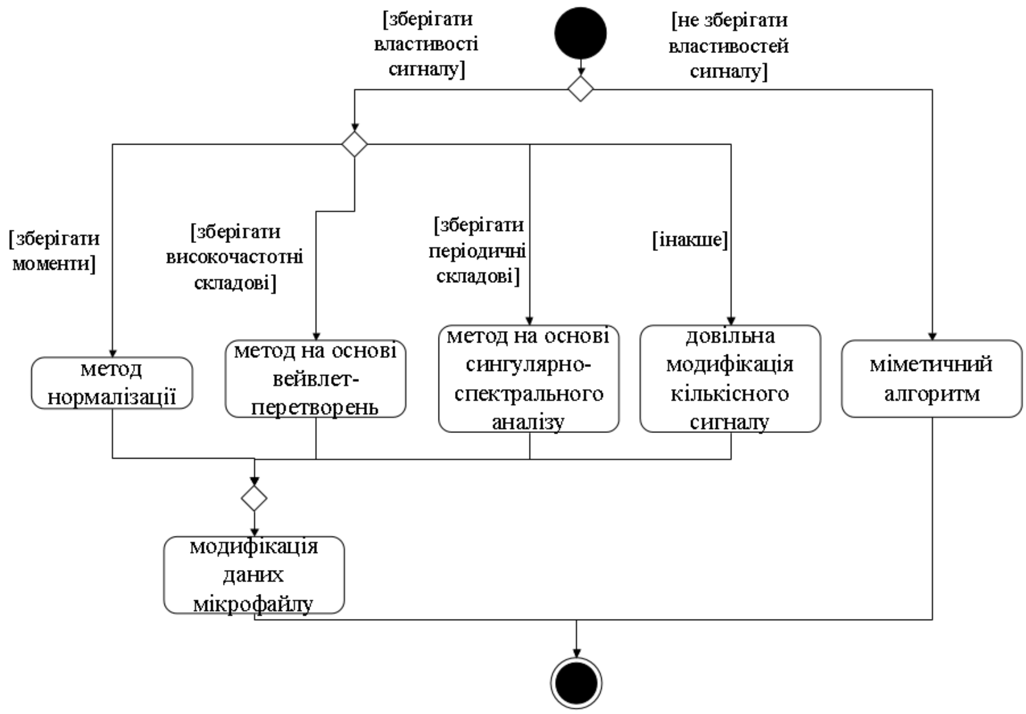


Рис. 2. UML-діаграма діяльності працівника організації-розпорядника даних, в обов’язки якого входить забезпечення групової анонімності даних

Висновки

У роботі виконано порівняльний аналіз описаних у літературі методів забезпечення групової анонімності даних як складової частини підпроцесу розповсюдження даних CDIS-процесу обробки даних. За результатами виконаного аналізу встановлено, що двоетапні методи забезпечення групової анонімності доцільно використовувати, коли особливості кількісного сигналу мають істотне значення для потенційних дослідників, і їх потрібно зберегти, навіть якщо обсяг спотворень, унесених у дані мікрофайлу, від цього збільшиться.

В умовах, коли структурні особливості кількісного сигналу не мають значної ваги, доцільно використовувати одноетапні методи забезпечення групової анонімності, зокрема, метод на основі міметичного алгоритму, який дає змогу одночасно модифікувати кількісний сигнал та дані мікрофайлу, уносячи, як правило, спотворення меншого обсягу.

Література

1. Duncan G.T. Exploring the tension between privacy and the social benefits of governmental databases / G.T. Duncan // Little Knowledge: Privacy, Security, and Public Information after September 11 [eds. J. Podesta, P.M. Shane, R.C.A. Leone]. — The Century Foundation : New York, 2004. — P. 71–88.
2. Duncan G.T. Statistical Confidentiality. Principles and Practice / G.T. Duncan, M. Elliot, G.J.J. Salazar. — New York : Springer-Verlag, 2011. — 212 p.
3. Gehrke J. Privacy in Data Publishing / J. Gehrke, D. Kifer, A. Machanavajjhala // IEEE 29th International Conference on Data Engineering (ICDE). — 2010. — P. 1213.
4. Little R.J.A. Statistical Analysis of Masked Data / R.J.A. Little // Journal of Official Statistics. — 1993. — 9(2). — P. 407–426.
5. A Terminology for Talking About Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34 [Electronic resource] / A. Pfitzmann, M. Hansen. — 2010. — Mode of access: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml
6. Evfimievski A. Randomization in Privacy Preserving Data Mining / A. Evfimievski // ACM SIGKDD Explorations Newsletter. — ACM Press, 2002. — 4(2). — P. 43–48.

7. Sweeney L. k-anonymity: a model for protecting privacy / L. Sweeney // International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems. — World Scientific, 2002. — 10(5). — P. 557–570.
8. Fienberg S. Data Swapping: Variations on a Theme by Dalenius and Reiss / S. Fienberg, J. McIntyre // Journal of Official Statistics. — 2005. — Vol. 21(2). — P. 309–324.
9. Domingo-Ferrer J. A survey of inference control methods for privacy-preserving data mining / J. Domingo-Ferrer // Privacy-Preserving Data Mining: Models and Algorithms. — Springer, 2008. — P. 53–80.
10. Parmar A.A. Blocking based approach for classification rule hiding to preserve the privacy in database / A.A. Parmar, U.P. Rao, D.R. Patel // International Symposium on Computer Science and Society. — 2011. — P. 323–326.
11. Liu Q.A Privacy-Preserving Data Publishing Method for Multiple Numerical Sensitive Attributes via Clust. and Multi-Sens. Bucket. / Q. Liu, H. Shen, Y. Sang // Int. Symp. on Par. Arch., Alg. and Progr. — 2014. — P. 220–223.
12. Singh A. Privacy preserving techniques in social networks data publishing-a review / A. Singh, D. Bansal, S. Sofat // International Journal of Computer Applications. — 2014. — Vol. 87, No. 15. — P. 9–14.
13. Rashid A.H. Privacy-preserving data publishing: review / A.H. Rashid, N.B.M. Yasin // International Journal of Physical Sciences. — 2015. — Vol. 10(7). — P. 239–247.
14. Тавров Д.Ю. Двофазовий меметичний алгоритм забезпечення групової анонімності даних / Д.Ю. Тавров, О.Р. Чертов // Штучний інтелект. — 2015. — № 1–2. — С. 170–179.
15. Chertov O. Microfiles as a Potential Source of Confidential Information Leakage / O. Chertov, D. Tavrov // Intelligent Methods for Cyber Warfare [ed. R.R. Yager, M.Z. Reformat, N. Alajlan]. — Springer International Publishing Switzerland, 2015. — P. 87–114.
16. Chertov O. Memetic Algorithm for Solving the Task of Providing Group Anonymity / O. Chertov, D. Tavrov // Advance Trends in Soft Computing [ed. M. Jamshidi, V. Kreinovich, J. Kacprzyk]. — Springer International Publishing Switzerland, 2014. — P. 281–292.
17. Chertov O. Two-Phase Memetic Modifying Transformation for Solving the Task of Providing Group Anonymity / O. Chertov, D. Tavrov // Recent Developments and New Direction in Soft-Computing Foundations and Applications [eds. L.A. Zadeh, A.M. Abbasov, R.R. Yager, S.N. Shahbazova, M.Z. Reformat]. — Springer International Publishing Switzerland, 2016. — P. 239–253.
18. Kleinberg J. Algorithm Design / J. Kleinberg, E. Tardos. — Pearson, 2005. — 864 p.
19. Liu L. Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statistics-Preserving / L. Liu, J. Wang, J. Zhang // 2008 IEEE International Conference on Data Mining Workshops. — IEEE Computer Society Press, 2008. — P. 27–35.
20. Chertov O. Providing Group Anonymity Using Wavelet Transform / O. Chertov, D. Tavrov // Data Security and Security Data [ed. L.M. MacKinnon]. — Berlin, Heidelberg : Springer-Verlag, 2012. — P. 25–36.
21. Golyandina N. The “Caterpillar”-SSA Method for Analysis of Time Series with Missing Values / N. Golyandina, E. Osipov // Journal of Statistical Planning and Inference. — 2007. — Vol. 137(8). — P. 2642–2653.
22. Neri F. A Primer on Memetic Algorithms / F. Neri, C. Cotta // Handbook of Memetic Algorithms [eds. F. Neri, C. Cotta, P. Moscato]. — Berlin, Heidelberg : Springer-Verlag, 2012. — P. 43–52.

Literatura

1. Duncan G.T. Exploring the tension between privacy and the social benefits of governmental databases / G.T. Duncan // Little Knowledge: Privacy, Security, and Public Information after September 11 [eds. J. Podesta, P.M. Shane, R.C.A. Leone]. — The Century Foundation : New York, 2004. — P. 71–88.
2. Duncan G.T. Statistical Confidentiality. Principles and Practice / G.T. Duncan, M. Elliot, G.J.J. Salazar. — New York : Springer-Verlag, 2011. — 212 p.
3. Gehrke J. Privacy in Data Publishing / J. Gehrke, D. Kifer, A. Machanavajjhala // IEEE 29th International Conference on Data Engineering (ICDE). — 2010. — P. 1213.
4. Little R.J.A. Statistical Analysis of Masked Data / R.J.A. Little // Journal of Official Statistics. — 1993. — 9(2). — P. 407–426.
5. A Terminology for Talking About Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34 [Electronic resource] / A. Pfitzmann, M. Hansen. — 2010. — Mode of access: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml
6. Evfimievski A. Randomization in Privacy Preserving Data Mining / A. Evfimievski // ACM SIGKDD Explorations Newsletter. — ACM Press, 2002. — 4(2). — P. 43–48.
7. Sweeney L. k-anonymity: a model for protecting privacy / L. Sweeney // International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems. — World Scientific, 2002. — 10(5). — P. 557–570.
8. Fienberg S. Data Swapping: Variations on a Theme by Dalenius and Reiss / S. Fienberg, J. McIntyre // Journal of Official Statistics. — 2005. — Vol. 21(2). — P. 309–324.
9. Domingo-Ferrer J. A survey of inference control methods for privacy-preserving data mining / J. Domingo-Ferrer // Privacy-Preserving Data Mining: Models and Algorithms. — Springer, 2008. — P. 53–80.

10. Parmar A.A. Blocking based approach for classification rule hiding to preserve the privacy in database / A.A. Parmar, U.P. Rao, D.R. Patel // International Symposium on Computer Science and Society. — 2011. — P. 323–326.
11. Liu Q. A Privacy-Preserving Data Publishing Method for Multiple Numerical Sensitive Attributes via Clust. and Multi-Sens. Bucket. / Q. Liu, H. Shen, Y. Sang // Int. Symp. on Par. Arch., Alg. and Progr. — 2014. — P. 220–223.
12. Singh A. Privacy preserving techniques in social networks data publishing-a review / A. Singh, D. Bansal, S. Sofat // International Journal of Computer Applications. — 2014. — Vol. 87, No. 15. — P. 9–14.
13. Rashid A.H. Privacy-preserving data publishing: review / A.H. Rashid, N.B.M. Yasin // International Journal of Physical Sciences. — 2015. — Vol. 10(7). — P. 239–247.
14. Tavrov D.Y. Dvofazovyi memetychnyi alhorytm zabezpechennia hrupovoii anonimnosti danykh / D.Y. Tavrov, O.R. Chertov // Shtuchnyi Intelekt. — 2015. — № 1–2. — S. 170–179.
15. Chertov O. Microfiles as a Potential Source of Confidential Information Leakage / O. Chertov, D. Tavrov // Intelligent Methods for Cyber Warfare [ed. R.R. Yager, M.Z. Reformat, N. Alajlan]. — Springer International Publishing Switzerland, 2015. — P. 87–114.
16. Chertov O. Memetic Algorithm for Solving the Task of Providing Group Anonymity / O. Chertov, D. Tavrov // Advance Trends in Soft Computing [ed. M. Jamshidi, V. Kreinovich, J. Kacprzyk]. — Springer International Publishing Switzerland, 2014. — P. 281–292.
17. Chertov O. Two-Phase Memetic Modifying Transformation for Solving the Task of Providing Group Anonymity / O. Chertov, D. Tavrov // Recent Developments and New Direction in Soft-Computing Foundations and Applications [eds. L.A. Zadeh, A.M. Abbasov, R.R. Yager, S.N. Shahbazova, M.Z. Reformat]. — Springer International Publishing Switzerland, 2016. — P. 239–253.
18. Kleinberg J. Algorithm Design / J. Kleinberg, E. Tardos. — Pearson, 2005. — 864 p.
19. Liu L. Wavelet-Based Data Perturbation for Simultaneous Privacy-Preserving and Statistics-Preserving / L. Liu, J. Wang, J. Zhang // 2008 IEEE International Conference on Data Mining Workshops. — IEEE Computer Society Press, 2008. — P. 27–35.
20. Chertov O. Providing Group Anonymity Using Wavelet Transform / O. Chertov, D. Tavrov // Data Security and Security Data [ed. L.M. MacKinnon]. — Berlin, Heidelberg : Springer-Verlag, 2012. — P. 25–36.
21. Golyandina N. The "Caterpillar"-SSA Method for Analysis of Time Series with Missing Values / N. Golyandina, E. Osipov // Journal of Statistical Planning and Inference. — 2007. — Vol. 137(8). — P. 2642–2653.
22. Neri F. A Primer on Memetic Algorithms / F. Neri, C. Cotta // Handbook of Memetic Algorithms [eds. F. Neri, C. Cotta, P. Moscato]. — Berlin, Heidelberg : Springer-Verlag, 2012. — P. 43–52.

RESUME

O.R. Chertov, D.Y. Tavrov

Providing group anonymity as a part of CSID data process

With the advent of modern information technologies, it has become an almost ubiquitous practice to provide public access to primary non-aggregated data to facilitate various kinds of research. When publishing such data, two contradictory interests clash. Respondents are usually interested in protecting as much sensitive information about themselves as possible, while potential data users would like to get access to as much data as possible. The process of providing public access to the data, and at the same time ensuring that sensitive information is protected, is called data stewardship. Organizations responsible for this kind of job are known as data stewardship organizations.

Such organizations process data according to what is called CSID data process (capture, storage, integration, dissemination). For the data protection part, dissemination subprocess is the most important, because it is at this stage when the data are modified to mask sensitive information. The modification applied must preserve sufficiently high level of data utility as well.

In practice, modifications during the dissemination process involve providing data anonymity, which can be of either individual or group kind. Methods for providing group anonymity have been proposed in the literature in the recent years, but they haven't been systematically analyzed in the broader context.

In this paper, group anonymity methods are put in the context of CSID data process. Appropriate conditions are identified for choosing particular methods depending on which part of information it is necessary to preserve during data modification.

Надійшла до редакції 19.10.2017

